Robust estimation and forecasting of the long-term seasonal component of electricity spot prices

Jakub Nowotarski and Jakub Tomczyk and Rafal Weron

Wrocław University of Technology, Poland, Wrocław University of Technology, Poland, Wrocław University of Technology, Poland

11. November 2012

Robust estimation and forecasting of the long-term seasonal component of electricity spot prices

Jakub Nowotarski[a], Jakub Tomczyk[a], Rafał Weron[b,*]

[a]*Hugo Steinhaus Center, Institute of Mathematics and Computer Science,*
Wrocław University of Technology, 50-370 Wrocław, Poland
[b]*Institute of Organization and Management, Wrocław University of Technology, 50-370 Wrocław, Poland*

Abstract

When building stochastic models for electricity spot prices the problem of uttermost importance is the estimation and consequent forecasting of a component to deal with trends and seasonality in the data. While the short-term seasonal components (daily, weekly) are more regular and less important for valuation of typical power derivatives, the long-term seasonal components (LTSC; seasonal, annual) are much more difficult to tackle. Surprisingly, in many academic papers dealing with electricity spot price modeling the importance of the seasonal decomposition is neglected and the problem of forecasting it is not considered. With this paper we want to fill the gap and present a thorough study on estimation and forecasting of the LTSC of electricity spot prices. We consider a battery of models based on Fourier or wavelet decomposition combined with linear or exponential decay. We find that all considered wavelet-based models are significantly better in terms of forecasting spot prices up to a year ahead than all considered sine-based models. This result questions the validity and usefulness of stochastic models of spot electricity prices built on sinusoidal long-term seasonal components.

Keywords: Electricity spot price, Long-term seasonal component, Robust modeling, Forecasting, Wavelets.

1. Introduction

As pointed out by Trück et al. (2007) and Janczura et al. (2012), the first crucial step in defining a model for electricity spot price dynamics consists of finding an appropriate description of the seasonal pattern. In the standard approach to seasonal decomposition the electricity spot price series P_t is decomposed into the trend-cycle or long-term seasonal component (LTSC) T_t, the periodic short-term seasonal component (STSC) s_t and remaining variability, error, or stochastic component X_t either in an additive (i.e., $P_t = T_t + s_t + X_t$) or a multiplicative fashion (i.e., $P_t = T_t \cdot s_t \cdot X_t$; note that a multiplicative model for the prices is equivalent to an additive model for

[*] Corresponding author

Email addresses: jakub.nowotarski@gmail.com (Jakub Nowotarski), jstomczyk@gmail.com (Jakub Tomczyk), rafal.weron@pwr.wroc.pl (Rafał Weron)

the logarithms of prices). The long-term seasonal component of electricity spot prices has been treated in the energy economics literature in a number of ways including:

- piecewise constant functions or dummies, possibly combined with a linear trend (Bhanot, 2000; Fanone et al., 2012; Fleten et al., 2011; Gianfreda and Grossi, 2012; Haldrup et al., 2010; Haugom and Ullrich, 2012; Higgs and Worthington, 2008; Keles et al., 2012a; Knittel and Roberts, 2005; Lucia and Schwartz, 2002),
- sinusoidal functions or sums of sinusoidal functions of different frequencies (Benth et al., 2012; Bierbrauer et al., 2007; Cartea and Figueroa, 2005; De Jong, 2006; Erlwein et al., 2010; Geman and Roncoroni, 2006; Keles et al., 2012b; Lucia and Schwartz, 2002; Pilipovic, 1998; Seifert and Uhrig-Homburg, 2007; Weron, 2008),
- wavelets (Conejo et al., 2005; Janczura and Weron, 2010, 2012; Stevenson, 2001; Schlueter, 2010; Stevenson et al., 2006; Weron, 2006, 2009; Weron et al., 2004a,b) or other nonparametric smoothing techniques (Bordignon et al., 2012).

However, to our best knowledge, there are only very few papers where the forecasting of the LTSC is discussed and even fewer where it is actually performed and checked. Forecasting a piecewise constant or a sinusoidal LTSC is straightforward, but it is either conducted for very short-term time horizons (e.g., one day-ahead as in Erlwein et al., 2010) or not conducted at all, probably due to the poor predictive power of such models outside carefully chosen time intervals and datasets. On the other hand, forecasting a nonparametric seasonal component is particularly troublesome and some authors actually evaluate only the out-of-sample prediction of the stochastic part (Bordignon et al., 2012). With this paper we want to fill the gap and present a thorough empirical study on estimation and forecasting of the LTSC of electricity spot prices. We consider a battery of models based on Fourier or wavelet decomposition, including models commonly used in the energy finance literature and a number of new suggestions.

The paper is structured as follows. In Section 2 we briefly describe the six datasets used in this empirical study. In the following Section we review different procedures for deseasonalizing the data and estimating the LTSC. In Section 4 we first outline the simulation setup, then present in detail all seven model families. Tables 1-2 can be used as a reference guide to the coding of the 300 models tested in this study. The Section ends with the definitions of error measures used later in the text. In Section 5 we report the results of our empirical study. We first discuss the global performance (over all six forecasting horizons and all six datasets), then comment on the performance across the forecasting horizons and finally discuss the results of a multiple comparison procedure which provides detailed information on which models perform significantly worse or significantly better than other models. In Section 6 we wrap up the results and comment on alternative approaches.

2. The data

To make the analysis and the resulting conclusions as universal as possible, in this study we use mean daily (baseload) spot prices from six major power markets:

- New South Wales Electricity Market (NSW; Australia) from the period Jan 1, 2006 – Jul 31, 2011 (2038 daily observations);
- European Energy Exchange (EEX; Germany) from the period Jan 1, 2001 – Apr 12, 2011 (3754 obs.);
- Nord Pool (NP; Norway) from the period Jan 1, 2000 – Nov 13, 2008 (3240 obs.);
- New England Power Pool (NEP; United States) from the period Jan 1, 2001 – Apr 28th, 2011 (3770 obs.);
- New York Independent System Operator (NYISO; United States) from the period Jan 1, 2004 – Jan 31, 2011 (2588 obs.);
- Pennsylvania–New Jersey–Maryland Market (PJM; United States) from the period Jan 1, 2006 – Apr 28, 2011 (1944 obs.).

The datasets are plotted in Figure 1. The annual seasonality is generally irregular, if visible at all. Note that this makes the probably most popular method of modeling the seasonal component with sine and cosine functions highly questionable. On the other hand, as shown by Janczura and Weron (2010), the changes in electricity price dynamics can be quite well linked to changes in market fundamentals. For instance, the electricity price hike in 2005 was largely due to higher natural gas (NG) prices, see the EEX (observations 1600-1850), NEP (obs. 1600-1850) and NYISO (obs. 500-750) price series. In Europe, the fuel prices were pushed up by the decline in North Sea production and a cold winter of 2005/2006. The introduction of CO_2 emission costs in January 2005 added momentum (Benz and Trück, 2006). In the U.S., the NG prices doubled after hurricanes Katrina and Rita damaged production, processing and transportation infrastructure. This volatile period was followed by roughly 18 months of more moderate prices and the second ‘fuel bubble’, which started in September/October 2007 and ended in July/August 2008 with the burst of the ‘oil bubble’ (Hamilton, 2009), see the EEX (obs. 2400-2900), NEP (obs. 2400-2900) and NYISO (obs. 1300-1800) price series.

The more regular weekly periodicity cannot be seen too well at this time scale. However, if we increase the resolution – as in Figure 4 – the five weekdays vs. weekend pattern is better visible. The price spikes tend to dominate Figure 1 and are visible in all six cases. Yet there are significant differences in the intensity and severity of the spikes. The NSW market is evidently the most spiky, the EEX dataset is pretty volatile and even includes a few price drops with a negative mean daily system price (for a discussion see, e.g., Fanone et al., 2012), while the NYISO market is the least spiky.

3. Estimating the long-term seasonal component

We follow the ‘industry standard’ and represent the spot price P_t by a sum of two independent parts: a stochastic component X_t and a (predictable) trend-seasonal component f_t composed of a weekly periodic part s_t (i.e., a short-term seasonal component, STSC) and a long-term trend-seasonal component (LTSC) T_t, which represents the long-term non-periodic fuel price levels, the

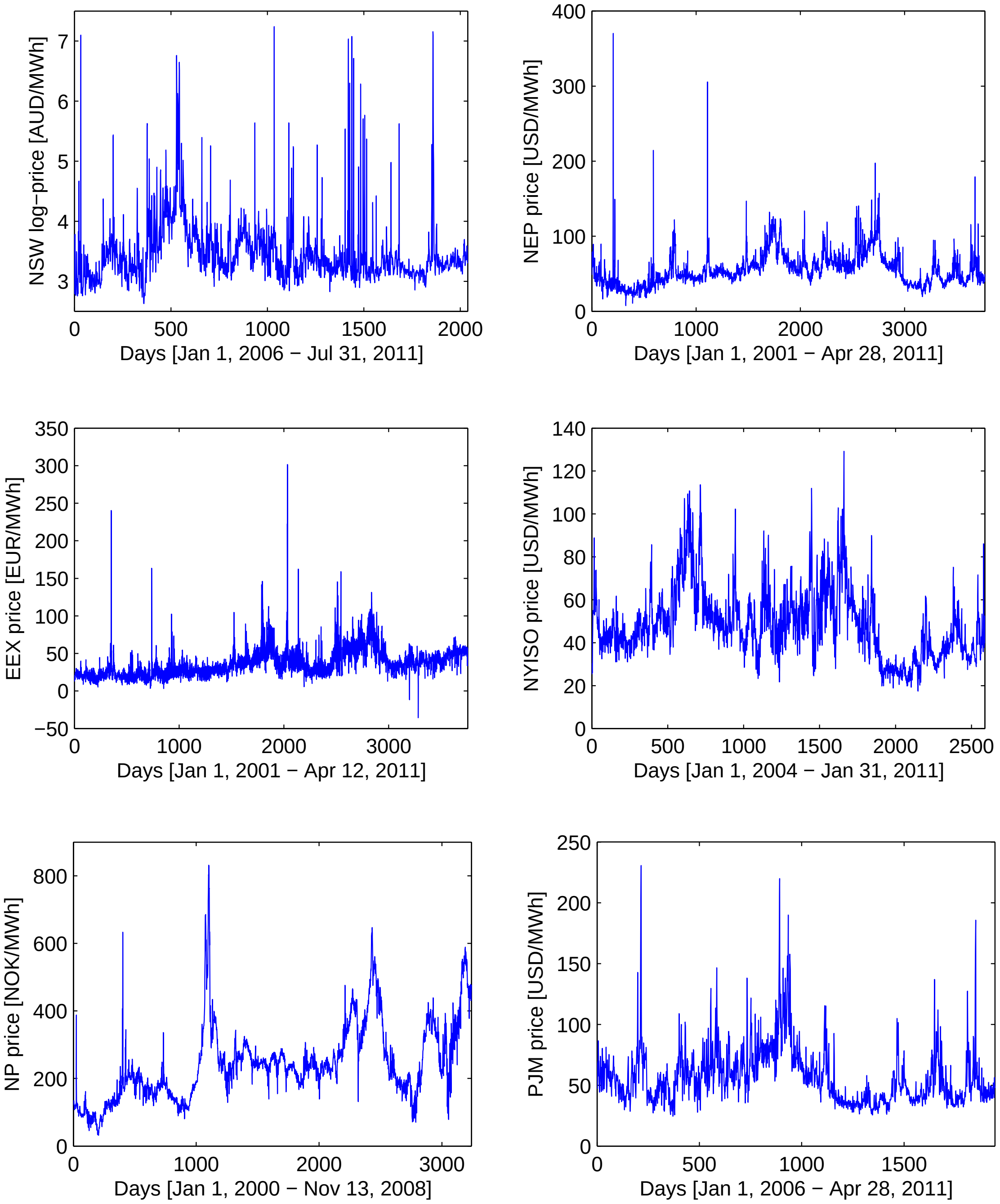

Figure 1: Mean daily (baseload) electricity spot prices from six major power markets (*from top to bottom, left to right*): New South Wales Electricity Market (NSW, Australia), European Energy Exchange (EEX, Germany), Nord Pool (NP, Scandinavia), New England Power Pool (NEP, U.S.), New York ISO (NYISO, U.S.) and Pennsylvania–New Jersey–Maryland Market (PJM, U.S.). Note that for the Australian market the log-prices (and not the prices themselves) are plotted. The logarithmic scale dampens the extreme spikiness of the NSW prices, which can reach up to 10000 AUD/MWh during peak hours.

changing climate/consumption conditions throughout the years and strategic bidding practices. As mentioned in the Introduction, there are essentially three distinct suggestions in the energy economics literature for dealing with the trend-seasonal component f_t of electricity spot prices (for a recent evaluation of the different approaches in terms of extracting the true seasonal pattern see Janczura et al., 2012). Sample fits of the LTSC are illustrated in Figure 2.

The first is to fit piecewise constant functions or dummies, typically one for each month. Forecasting of such a LTSC is trivial. However, given the not very periodic behavior of spot prices on the annual scale (recall Figure 1), the usefulness of this technique is questionable. Moreover, while very simple, this approach yields a non-smooth trend-seasonal component with jumps between months. If this effect is not eliminated by an additional smoothing treatment, it may very well introduce spurious seasonality and negatively influence the estimation of the stochastic component. In our study we will use as benchmarks only very simple variants of this technique which do not require additional smoothing, namely a constant and a linear LTSC.

The second approach is to model the trend-seasonal pattern T_t by a sum of sine and/or cosine functions. Due to the rather complex annual pattern of spot electricity prices, except for a few regular periods like the Jan 1997 – Apr 2000 period at Nord Pool analyzed by Weron (2008), the LTSC cannot be modeled by a single sine function. The question whether the periods of other sine or cosine functions of higher frequency should be harmonics of the annual frequency or not is an open one. In order to answer it we will consider here sinusoidal models for T_t with up to four summands with both regularly (harmonics) and irregularly spaced frequencies. Note that in both cases the forecasting of such a LTSC is straightforward, since it is based on a simple extrapolation of the sine and/or cosine functions with known frequency, phase shift and amplitude.

The third approach is to use wavelet decomposition and smoothing as more robust to outliers and a less periodic alternative to Fourier analysis. Recall, that wavelets belong to families – like the Daubechies and Coiflets families used here – and come in pairs of a father and a mother wavelet for a given order (Härdle et al., 1998; Percival and Walden, 2000). The different families and orders of wavelets make different trade-offs between how compactly they are localized in time and their smoothness. Any function or signal (here, P_t) can be built up as a sequence of projections onto one father wavelet and a sequence of mother wavelets: $S_J + D_J + D_{J-1} + ... + D_1$, where 2^J is the maximum scale sustainable by the number of observations. At the coarsest scale the signal can be estimated by S_J. At a higher level of refinement the signal can be approximated by $S_{J-1} = S_J + D_J$. At each step, by adding a mother wavelet D_j of a lower scale $j = J-1, J-2, ...$, we obtain a better estimate of the original signal. This procedure, also known as lowpass filtering, yields a traditional linear smoother. Here we use $J = 6$, 7 and 8, which roughly correspond to bimonthly ($2^6 = 64$ days), seasonal ($2^7 = 128$ days) and annual ($2^8 = 256$ days) smoothing. While trigonometric or periodic functions – such as the sinusoidal LTSC or the monthly dummies – can be easily extrapolated into the future, predicting the wavelet LTSC beyond the next few weeks is a difficult task. This results from the fact that, in contrast to sines or cosines, individual wavelet functions are quite localized in time or (more generally) in space. In the following Section we will address this issue and propose a few solutions. Note that also combining sinusoidal functions with an exponentially weighted moving average (as in De Jong, 2006) complicates things very much, because the moving average at time $t + 1$ is dependent on the unknown future price P_{t+1}.

The next and final step of seasonal decomposition would be to remove the weekly periodicity

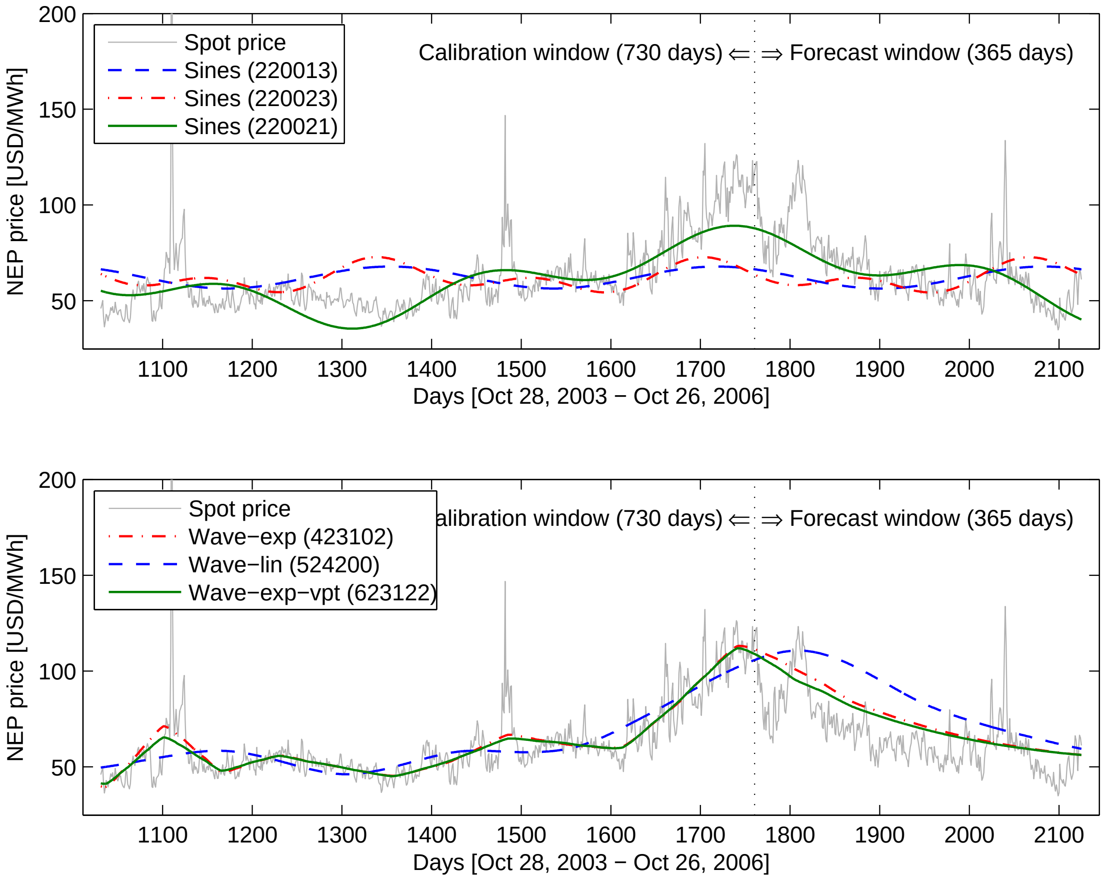

Figure 2: Sample LTSC estimation and forecasting results for the NEPOOL market and a two-year calibration window Oct 28, 2003 – Oct 26, 2005. Clearly the sinusoidal LTSC (*top panel*) do not follow the trend-seasonal pattern of the spot prices in the calibration window as well as the wavelet LTSC (*bottom panel*), even if the periods – as well as the amplitudes and phases – of the sine waves are estimated within an optimization procedure (model 220021; for code definitions see Section 4.2 and Tables 1-2).

s_t, typically by subtracting the 'average week' calculated as the arithmetic mean of the LTSC-deseasonalized prices (i.e., $P_t - T_t$) corresponding to each day of the week, with public holidays treated as the eighth day of the week. However, we will not explicitly estimate the weekly periodicity here, since for the time horizons considered in this study (see below) it is of little importance. Finally, note that the forecasting of such a periodic STSC is straightforward, like in the case of a LTSC build on piecewise constant functions (or dummies).

4. Forecasting the long-term seasonal component

4.1. The simulation setup

In the simulation study we use a rolling window scheme. At each estimation/forecasting step both the starting and the ending date of the calibration sample is moved forward by one day. Each of the 300 models we consider is estimated on two *calibration windows*: a two-year (730-day) and a three-year (1095-day). The following year (365 days) is used for the out-of-sample forecast and is denoted in the text as the *forecast window*, see Figures 2-4. In order to have the same number of forecasts for both calibration windows, the two-year window has a 365-day lag with respect to the three-year window (i.e., it starts on the 366th observation of the three-year calibration window). The rolling scheme lasts as long as we have at least 365 observations following the last day of the calibration window. In this way we obtain 579 forecasts (from 1 to 365 days ahead) for the NSW market, 2295 for the German EEX market, 1781 for the Scandinavian Nord Pool power exchange, 2311 for the New England Power Pool, 1129 for the New York ISO and 485 for the PJM market. After computing forecasts for all six datasets, all 300 models and all calibration windows we calculate three measures of forecast accuracy – the mean absolute error (MAE), the mean squared error (MSE) and the mean absolute percentage error (MAPE) – for each dataset, each model and each of the six forecast horizons: 1-7 days, 8-30 days, 31-90 days, 91-182 days, 183-274 days and 275-365 days.

4.2. Models and their codes

To cope with the large number of models used in this study each model is given a unique 6-digit code, see Tables 1 and 2. The first digit defines the model family, the second provides information on the calibration window (two-year – ‘2’, three-year –‘3’). The remaining four digits define family-specific characteristics. A star (‘*’) indicates that a certain digit can take one of a few values and is used to represent subgroups of models.

*4.2.1. Simple models (1*000*)*

In Figures 2-4 we can observe that when 2, 3 or 4 sines are fitted to raw or spike-filtered spot price data (models 2***00 and models 3****0, respectively; see Section 4.2.2 for details) the price forecast tends to deviate significantly from a reasonable price range, especially for longer time horizons. This is the reason for using conservative, simple techniques in this study. Initially we started with three models:

- the mean of the spot price in the calibration window (models **1*0001**),
- linear regression of the spot price in the calibration window extrapolated into the forecasting window (models **1*0002**) and
- the median of the spot price in the calibration window (models **1*0003**).

All three models performed surprisingly well for long-term forecasts of six months or more and the best of the three was the median. However, the short- and medium-term forecasts were significantly worse than those of the other models, mainly due to the price spikes and the heteroskedasticity of the spot price. Hence, we decided to test three more simple models:

- an exponential decay from the current spot price to the median (model **1*0004** with the decay parameter $\lambda = \frac{1}{30}$ in formula (1) and model **1*0005** with $\lambda = \frac{1}{180}$) and
- a linear decay from the current spot price to the median (model **1*0006**).

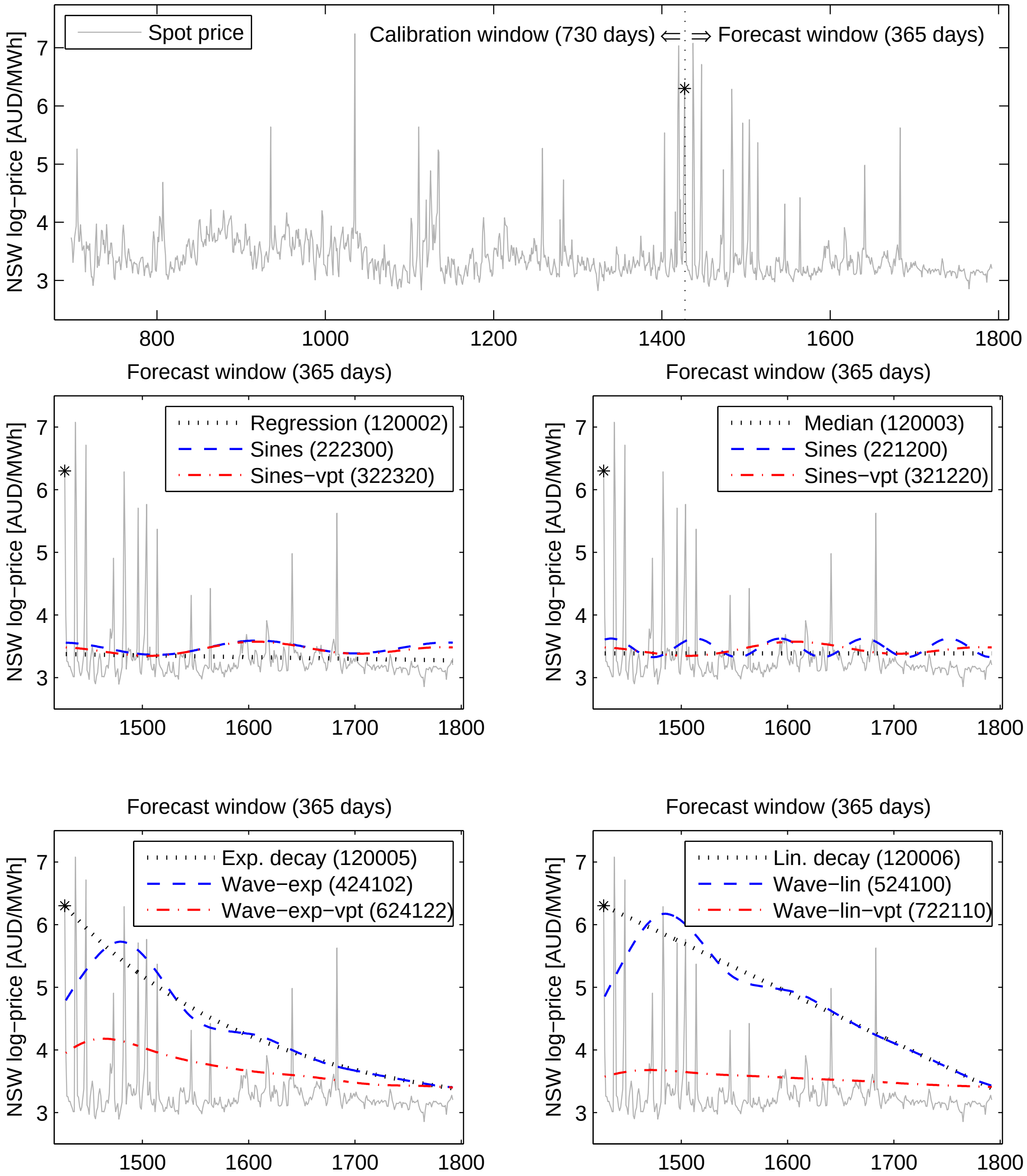

Figure 3: Sample LTSC forecasting results for the Australian NSW market performed on Friday, Nov 27, 2009 (denoted by '*'). The two-year calibration window and the one-year forecast window are displayed in the *top panel*. 12 different forecasting methods are illustrated in the *middle* and *bottom panels* zooming in on the forecast day (Nov 27, 2009) and the forecast window (Nov 28, 2009 – Nov 27, 2010). For code definitions see Section 4.2 and Tables 1-2.

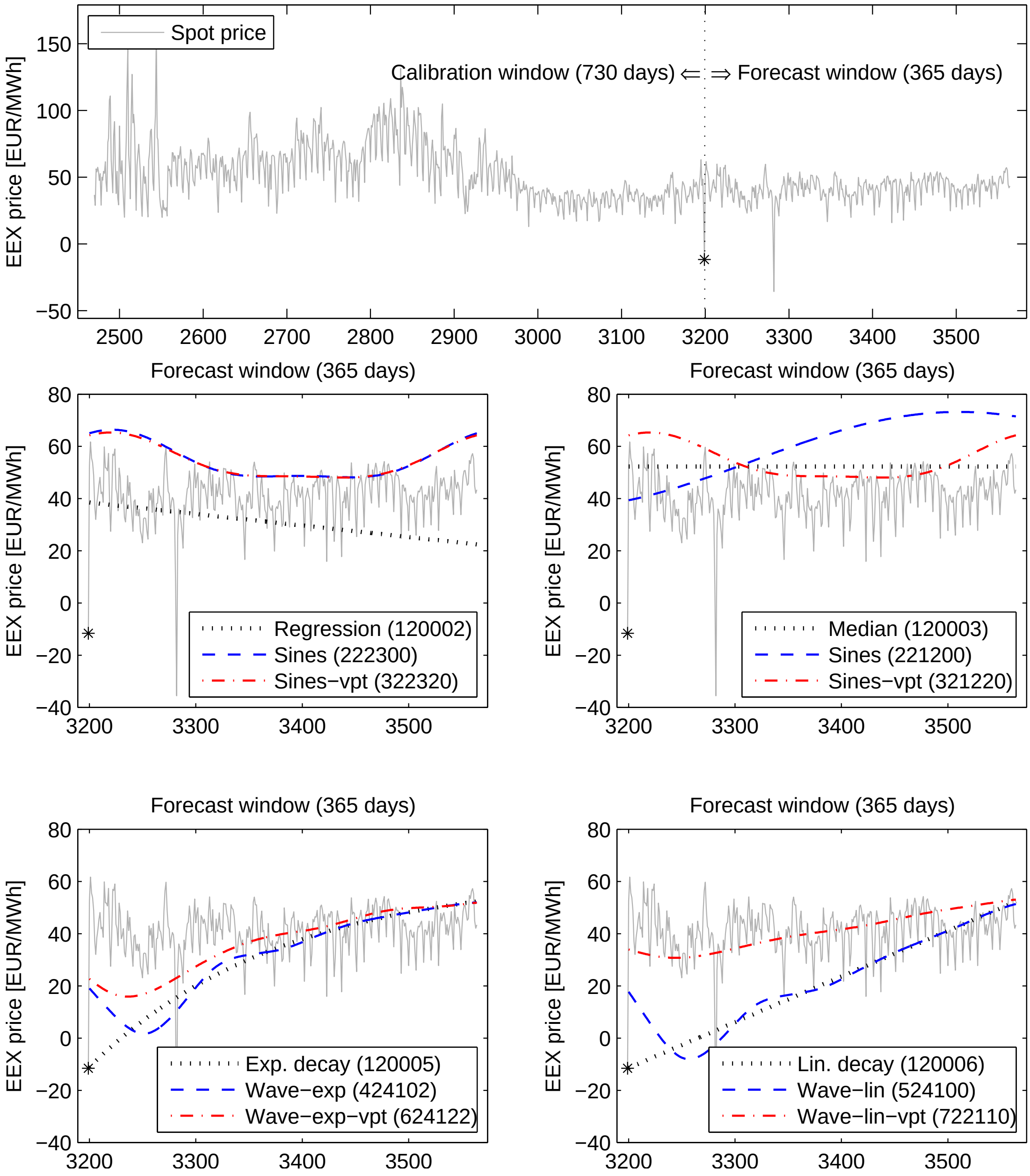

Figure 4: Sample LTSC forecasting results for the Nord Pool market performed on Sunday, Oct 4, 2009 (denoted by '*'). The two-year calibration window and the one-year forecast window are displayed in the *top panel*. 12 different forecasting methods are illustrated in the *lower panels* zooming in on the forecast day (Oct 4, 2009) and the forecast window (Oct 5, 2009 – Oct 4, 2010). For code definitions see Section 4.2 and Tables 1-2.

Table 1: The six digit codes of the 300 models tested in this study, part I. A star (‘*’) indicates that a certain digit can take one of a few values and is used to represent subgroups of models. A square cup (‘⊔’) identifies the digit of interest.

Digit	Value	Meaning
All models		
*⊔****	2	two year (730 day) calibration window
	3	three year (1095 day) calibration window
Simple models (1*000*) → 12 models in total		
1*000⊔	1	mean price in the calibration window
	2	extrapolated linear regression of prices in the calibration window
	3	median price in the calibration window
	4	exponential decay to the median with the decay parameter $\lambda = \frac{1}{30}$
	5	exponential decay to the median with the decay parameter $\lambda = \frac{1}{180}$
	6	linear decay to the median
Sines fitted to raw prices (2***00) → 24 models		
2*⊔*00	1,...,4	number of sines used to represent the LTSC
2**⊔00	1	periods of all sines estimated
	2	period of the 1st sine estimated, remaining periods set to 1, $\frac{1}{2}$ and $\frac{1}{3}$ of a year
	3	periods set to 1, $\frac{1}{2}$, $\frac{1}{3}$ and $\frac{1}{4}$ of a year
Sines fitted to spike-filtered prices (3****0) → 48 models		
3*⊔**0	1,...,4	number of sines used to represent the LTSC
3**⊔*0	1	periods of all sines estimated
	2	period of the 1st sine estimated, remaining periods set to 1, $\frac{1}{2}$ and $\frac{1}{3}$ of a year
	3	periods set to 1, $\frac{1}{2}$, $\frac{1}{3}$ and $\frac{1}{4}$ of a year
3***⊔0	1	spikes replaced by the mean of the deseasonalized prices
	2	spikes replaced by the upper/lower 2.5% quantiles of the deseasonalized prices
Wavelets with an exponential decay to the median fitted to raw prices (4***0*) → 48 models		
4*⊔*0*	1	Daubechies wavelet family of order 12 (‘db12’)
	2	Daubechies wavelet family of order 24 (‘db24’)
	3	Coiflets wavelet family of order 2 (‘coif2’)
	4	Coiflets wavelet family of order 4 (‘coif4’)
4**⊔0*	1	S_6 approximation level
	2	S_7 approximation level
	3	S_8 approximation level
4***0⊔	1	exponential decay to the median with the decay parameter $\lambda = \frac{1}{30}$
	2	exponential decay to the median with the decay parameter $\lambda = \frac{1}{180}$
Wavelets with a linear decay to the median fitted to raw prices (5***00) → 24 models		
5*⊔*00	1	Daubechies wavelet family of order 12 (‘db12’)
	2	Daubechies wavelet family of order 24 (‘db24’)
	3	Coiflets wavelet family of order 2 (‘coif2’)
	4	Coiflets wavelet family of order 4 (‘coif4’)
5**⊔00	1	S_6 approximation level
	2	S_7 approximation level
	3	S_8 approximation level

Table 2: The six digit codes of the 300 models tested in this study, part II.

Digit	Value	Meaning
Wavelets with an exponential decay to the median fitted to spike-filtered prices (6*****) → 96 models		
6*⊔***	1	Daubechies wavelet family of order 12 (‘db12’)
	2	Daubechies wavelet family of order 24 (‘db24’)
	3	Coiflets wavelet family of order 2 (‘coif2’)
	4	Coiflets wavelet family of order 4 (‘coif4’)
6**⊔**	1	S_6 approximation level
	2	S_7 approximation level
	3	S_8 approximation level
6***⊔*	1	spikes replaced by the mean of the deseasonalized prices
	2	spikes replaced by the upper/lower 2.5% quantiles of the deseasonalized prices
6****⊔	1	exponential decay to the median with the decay parameter $\lambda = \frac{1}{30}$
	2	exponential decay to the median with the decay parameter $\lambda = \frac{1}{180}$
Wavelets with an exponential decay to the median fitted to spike-filtered prices (7****0) → 48 models		
7*⊔**0	1	Daubechies wavelet family of order 12 (‘db12’)
	2	Daubechies wavelet family of order 24 (‘db24’)
	3	Coiflets wavelet family of order 2 (‘coif2’)
	4	Coiflets wavelet family of order 4 (‘coif4’)
7**⊔*0	1	S_6 approximation level
	2	S_7 approximation level
	3	S_8 approximation level
7***⊔0	1	spikes replaced by the mean of the deseasonalized prices
	2	spikes replaced by the upper/lower 2.5% quantiles of the deseasonalized prices

All three models connect the last day of the calibration period – the current spot price – and the last day of the forecast window – the median of the spot prices in the calibration window. The exponential decay function is normalized in the following way:

$$f_{exp}(x) = (x_0 - x_T) \times \frac{\exp(-\lambda x) - \exp(-\lambda x_T)}{1 - \exp(-\lambda x_T)} + x_T, \tag{1}$$

where x_0 is the last observation in the calibration window (i.e., on the day the prediction in made), x_T is the median of the spot prices in the calibration window (on the last day of the forecast window) and $\lambda = \frac{1}{30}$ or $\frac{1}{180}$. Since $\exp(-\lambda x)$ is the tail of the exponential distribution with mean $\frac{1}{\lambda}$, the forecast of model 1*0004 decays to the median much faster (i.e., its mean lifetime is 30 days or one month) than that of model 1*0005 (whose mean lifetime is 180 days or half a year).

*4.2.2. Sines fitted to raw (2***00) or spike-filtered prices (3****0)*

The second family of models considered in this study and denoted by 2***00 used sine functions to represent the LTSC. In these models a sum of one to four sines is fitted via nonlinear least squares to the spot price in the calibration window – the third digit in the model code (**2*⊔*00**) stands for the number of sines used to represent the LTSC (1, 2, 3 or 4). Each considered sine function has three parameters to estimate – the amplitude, the period and the phase shift. To address the question whether the periods of the sine functions of higher frequency should be harmonics of the annual frequency or not we consider three subgroups of models:

- models where the periods of all sines (up to four) are estimated within the least squares procedure (models **2**100**, i.e., with ‘1’ as the fourth digit),
- models where the period of the first sine function is estimated within the least squares procedure and the periods of the remaining sine functions (if any) are set to a year, half a year and a third of a year (models **2**200**, i.e., with ‘2’ as the fourth digit) and
- models where the periods are set to a year, half a year, a third of a year and a quarter of a year (models **2**300**, i.e., with ‘3’ as the fourth digit).

Note that in all three cases the forecasting of such a LTSC is straightforward, since it is based on a simple extrapolation of the sine functions with known frequency, phase shift and amplitude. Note also that the latter method should not outperform the other two; even if the fixed periods were optimal then they should be estimated within the least squares procedure for the first two methods. Yet, the obtained forecasting results do not match our expectations. Most likely, the nonlinear least squares optimization procedure has problems with finding the global maximum due to the large number of parameters to be estimated (up to 12).

In a recent empirical study Janczura et al. (2012) showed that improved robustness of the electricity spot price model could be achieved by filtering the data with some reasonable procedure for outlier (i.e., spike) detection, and then using classical estimation techniques for the seasonal pattern on the filtered data. While no single best method for outlier detection could be identified, in a vast majority of cases all of the considered filtering techniques significantly outperformed the ‘no filter’ approach that used the original spot price. Out of the seven filtering techniques tested in this study, the simple-to-implement *2.5% variable price thresholds* (VPT1) method yielded reasonable improvement over the ‘no filter’ approach, both with respect to estimating the seasonal pattern (LTSC and STSC) and the parameters of the stochastic component. In this method 2.5% highest and 2.5% lowest deseasonalized prices are treated as outliers and are replaced by ‘more normal’ values. The deseasonalization is performed by first subtracting a wavelet smoother of level 6 (i.e., S_6, see Section 3; or a sine function combined with an exponentially weighted moving average) from the spot prices, then by computing the ‘average week’ (with holidays treated as the eighth day of the week) and removing it from the LTSC-detrended data. Janczura et al. (2012) used the seasonal pattern as the ‘more normal’ values – they substituted the identified spikes in the deseasonalized series by the mean of the deseasonalized prices.

We have decided to use this technique in our study. The sine-based models fitted to spike-filtered prices constitute the third family of models (3****0). The third and the fourth digit in the model code define the same characteristics as in the standard sine-based models (2***00). The fifth digit defines the ‘more normal’ values used to substitute the identified spikes:

- models **3***10** are fitted to price series where the identified spikes a replaced by the mean of the deseasonalized prices, as in Janczura et al. (2012),
- whereas models **3***20** are fitted to price series where the identified spikes are replaced by the threshold itself (i.e., the upper or lower 2.5% quantile of the deseasonalized prices).

Note that the latter approach is similar in spirit to the *limiting* spike preprocessing scheme used in the engineering literature (Shahidehpour et al., 2002; Weron, 2006).

*4.2.3. Wavelets (4***0*, 5***00, 6*****, 7****0)*

The next four families of models we consider consist of wavelet-based LTSC. They differ in the way the signal is extrapolated before applying the Discrete Wavelet Transform (DWT) and the choice of the input signal (raw or spike-filtered prices). Since the lengths of the two- and the three-year calibration windows are not powers of two, the estimation procedure requires that the signal is artificially extended before applying the DWT, so that its length is the nearest power of two (Härdle et al., 1998; Percival and Walden, 2000). However, the commonly used extension modes, like constant extension (in Matlab denoted by 'sp0') and first derivative extrapolation ('sp1') at the edges, do not perform satisfactorily in case of spiky electricity prices. Other, more appropriate techniques have to be applied.

On the other hand, a characteristic feature of wavelets is that – unlike sines and cosines – the individual wavelet functions are localized in time, i.e., they tend to zero for large (positive or negative) arguments. In the context of extrapolating the signal into the future this means that some additional assumptions have to be made on how to extrapolate the smoothed signal S_j and/or the detail series of lower orders. There have been a few suggestions in the literature on how to deal with this problem. For the lower detail levels, which are of high frequency and oscillatory in nature, Yousefi et al. (2005) used trigonometric fits in a study of oil prices while Conejo et al. (2005) calibrated ARIMA models when forecasting day-ahead electricity spot prices in the Spanish market. Fryźlewicz et al. (2003) proposed yet a more technical concept – the locally stationary wavelet process, where the price process is written as a linear combination using wavelets as basis functions. Although appealing from the theoretical point of view, it performed very poorly in a recently published short-term forecasting study of Schlueter and Deuschle (2010). In fact, the authors observed that for signals with a strong random component – like the UK power prices – all tested wavelet-based methods at best generated only little improvements over the more traditional time series approaches.

For the much smoother and less periodic approximations (like the S_6, S_7 or S_8 approximations used in this study) and the higher detail levels other techniques have to be applied. To extend the signal beyond the calibration window, Yousefi et al. (2005) used a spline fit, Wong et al. (2003) applied polynomial extrapolation, while Stevenson (2001) and Stevenson et al. (2006) utilized predictions of threshold autoregressive models (TAR) fitted to smoothed (via wavelet shrinkage) spot prices from the Australian electricity market. To our best knowledge, the latter two papers are the only ones where wavelets have been used to forecast electricity prices for horizons of more than a few days ahead. Yet both papers used prices from the out-of-sample period to extend the signal to the nearest power of two and avoid edge extension problems; hence the predictions were not truly *ex-ante* forecasts. This is a pretty controversial approach which clearly cannot be used in real world applications.

Taking into account that in this study we are only interested in *ex-ante* forecasts with a relatively long time horizon, none of the methods mentioned above could be applied; the spline- or polynomial extrapolation-based forecasts of electricity spot prices behaved unpredictably over periods of a few hundred days. Hence, we decided to extrapolate the smoothed signal similarly as in the case of sine-based models, i.e., by fitting a sum of up to four sines to S_j in the calibration window and simply extrapolating the sines 365 days into the future. Unfortunately, this too resulted in unreasonably fluctuating predictions of future electricity spot prices. In the next attempt – motivated by the surprisingly good long-term forecasts of the simple models (recall the discussion in Section 4.2.1) and the relatively good short-term forecasts of all wavelet models compared to the simple (1*000*) and sine-based models (2***00, 3****0) – we introduced two new families:

- wavelets with an exponential decay to the median (**4***0***) and
- wavelets with a linear decay to the median (**5***00**).

Two further families of models are their analogues fitted to spike-filtered prices: **6******* and **7****0**, respectively. Like for the sine-based models (i.e., 3****0), the fifth digit defines the 'more normal' values used to substitute the identified spikes:

- models **6***1*** and **7***10** are fitted to price series where the identified spikes a replaced by the mean of the deseasonalized prices, as in Janczura et al. (2012),
- whereas models **6***2*** and **7***20** are fitted to price series where the identified spikes are replaced by the threshold itself (i.e., the upper or lower 2.5% quantile of the deseasonalized prices).

Instead of using the 'sp0' or 'sp1' extension modes at the edges, in these new models the calibration windows are first extended one year forward using an exponentially or a linearly decaying to the median deterministic function. This is done analogously as in models 1*0004, 1*0005 and 1*0006: we connect the last day of the calibration period (on the time axis) and the current spot price (on the price axis) with the last day of the forecast window (on the time axis; observation 1095 for the two-year calibration window and observation 1460 for the three-year calibration window) and the median of the spot prices in the calibration window (on the price axis). The exponential decay function is given by formula (1), either with $\lambda = \frac{1}{30}$ (models **4***01** and **6****1**) or $\frac{1}{180}$ (models **4***02** and **6****2**). Once the data series are extended to 1095 (or 1460) observations we apply the DWT. Note that the wavelet estimation procedure again has to extend the series so that its length is a power of two – this time to 2048 observations for both calibration windows. However, now the constant extension at the edges (i.e., 'sp0') does not influence the shape of the wavelet smoother too much since the last observation of the initially extended series (of 1095 or 1460 observations) is the median. Finally, we simply take as our forecast the 365 values of the obtained wavelet smoother corresponding to the forecast window.

For each wavelet-based model family we use four types of wavelets which make different trade-offs between how compactly they are localized in time and their smoothness: two from the Daubechies family (of order 12 and 24; in Matlab and later in the text denoted by 'db12' and 'db24', respectively) and two from the Coiflets family (of order 2 and 4; denoted by 'coif2' and 'coif4', respectively). The third digit in the model code (****⊔*****) defines the wavelet: '1' stands for 'db12', '2' for 'db24', '3' for 'coif2' and '4' for 'coif4'. Finally, for each wavelet-based family the fourth digit in the model code (*****⊔****) defines the wavelet approximation level: '1' stands for S_6, '2' for S_7 and '3' for S_8. Note that the three approximation levels used roughly correspond to bimonthly ($2^6 = 64$ days), seasonal ($2^7 = 128$ days) and annual ($2^8 = 256$ days) smoothing, respectively.

4.3. Error measures

Six datasets, over 17 thousand observations and as many as 300 models tested lead us to a fundamental question: how to select the best LTSC forecasting technique(s)? To address this question we calculate three measures of forecast accuracy – the mean absolute error (MAE), the mean squared error (MSE) and the mean absolute percentage error (MAPE) – for each dataset, each model and each forecast horizon. Then we rank the models from 1 to 300 based on the values of $\text{MAE}_{h,d}$ or $\text{MSE}_{h,d}$:

- separately for each of the six forecast horizons: $h = 1$ (1-7 days), 2 (8-30 days), 3 (31-90 days), 4 (91-182 days), 5 (183-274 days) and 6 (275-365 days) and
- each of the six datasets $d = 1$ (NSW), 2 (EEX), 3 (NP), 4 (NEP), 5 (NYISO) and 6 (PJM).

To obtain the aggregate rank (over all six datasets) of a model for a given time horizon h we calculate the geometric mean – denoted by $\text{GM}(\text{MAE}_{h,*})$ or $\text{GM}(\text{MSE}_{h,*})$ – of the six ranks for each of the six datasets for this time horizon. Note that compared to the arithmetic mean, the geometric mean penalizes poor rankings and emphasizes good rankings. Next, we compute an average rank over all time horizons: we rank the models from 1 to 300 based on the values of $\text{GM}(\text{MAE}_{h,*})$ or $\text{GM}(\text{MSE}_{h,*})$ for each of the six time horizons and compute the geometric mean of those six ranks. The resulting two global measures are denoted by $\text{GM}(\text{MAE}_{*,*})$ and $\text{GM}(\text{MSE}_{*,*})$.

Furthermore, since the ranks do not provide quantitative information about a given method's forecasting accuracy we use two aggregate measures based on the individual mean absolute percentage errors ($\text{MAPE}_{h,d}$ with $h, d = 1, ..., 6$). Namely, for a given time horizon h we calculate the weighted arithmetic mean

$$\text{MAPE}_{h,*} = \sum_{d=1}^{6} w_d \cdot \text{MAPE}_{h,d}, \tag{2}$$

where $w = \left(\frac{579}{8580}, \frac{2295}{8580}, \frac{1781}{8580}, \frac{2311}{8580}, \frac{1129}{8580}, \frac{485}{8580}\right)$ is the vector of weights such that each dataset has a weight proportional to its length. Next, we compute the average over all time horizons:

$$\text{MAPE}_{*,*} = \sum_{h=1}^{6} v_h \cdot \text{MAPE}_{h,*}, \tag{3}$$

where $v = \left(\frac{7}{365}, \frac{23}{365}, \frac{60}{365}, \frac{92}{365}, \frac{92}{365}, \frac{91}{365}\right)$ is the vector of weights such that each forecasting horizon has a weight proportional to its length. For a given model the global error measure $\text{MAPE}_{*,*}$ is the mean absolute percentage error over all datasets and all forecasting horizons.

5. Results

5.1. Global performance

In Table 3 we list the top 20 models according to each of the three global forecast error measures: $\text{GM}(\text{MAE}_{*,*})$, $\text{GM}(\text{MSE}_{*,*})$ and $\text{MAPE}_{*,*}$. Nearly all models in the top 20 list are from

Table 3: Top 20 models according to each of the three global forecast error measures: GM(MAE$_{*,*}$) in columns 2-3, GM(MSE$_{*,*}$) in columns 4-5 and MAPE$_{*,*}$ in columns 6-7. The best three models in terms of each measure are emphasized in bold with the index indicating their rank: ^{a,b,c} for GM(MAE$_{*,*}$), ^{1,2,3} for GM(MSE$_{*,*}$) and ^{A,B,C} for MAPE$_{*,*}$. Note that all models in the top 20 list are from families 6***** and 7****0; the best models from the remaining five families are listed in the bottom rows of the table. See Section 4.2 for model codes and Section 5 for error measure definitions.

No.	GM(MAE$_{*,*}$)	Model	GM(MSE$_{*,*}$)	Model	MAPE$_{*,*}$	Model
1	14.58	**733110**^{a,B}	14.19	**622322**^{1}	29.36%	**734110**^{c,A}
2	15.40	**731310**^{b,E}	15.02	**624322**^{2}	29.38%	**733110**^{a,B}
3	20.48	**734110**^{c,A}	16.73	**623322**^{3}	29.41%	**732110**^{e,C}
4	21.00	**633122**^{d}	17.91	**631312**^{4}	29.44%	**731110**^{D}
5	22.26	**732110**^{e,C}	18.07	**631322**^{5}	29.57%	**731310**^{b,E}
6	22.59	**631312**^{4}	27.24	**633122**^{d}	29.73%	734210
7	22.97	634122	28.05	621322	29.74%	733210
8	27.88	734120	28.47	634122	29.76%	732310
9	29.37	722110	28.98	624122	29.77%	731320
10	30.23	**631322**^{5}	33.70	621122	29.78%	734120
11	32.40	631222	33.95	**731310**^{b,E}	29.80%	734310
12	33.45	733120	34.45	721120	29.82%	731120
13	33.66	**623322**^{3}	34.68	623122	29.84%	733120
14	33.94	**731110**^{D}	34.71	722320	29.84%	732210
15	33.96	721110	35.58	632322	29.85%	732320
16	34.04	623122	35.86	431302	29.86%	732120
17	34.73	731120	36.48	722220	29.86%	731210
18	35.90	723320	37.24	424302	29.87%	**631322**^{5}
19	36.73	624122	37.52	724320	29.90%	633112
20	36.98	724110	38.50	734120	29.91%	734320
⋮	⋮	⋮	⋮	⋮	⋮	⋮
55	55.11	431302	.	.	.	.
68	.	.	.	.	30.28%	431302
72	.	.	.	.	30.35%	531300
79	.	.	73.15	524200	.	.
80	68.95	521300	.	.	.	.
87	71.75	120005	.	.	.	.
110	.	.	90.89	120005	.	.
137	.	.	.	.	31.21%	120005
201	.	.	162.73	333110	.	.
203	.	.	163.29	231110	.	.
221	191.50	333110	.	.	.	.
223	.	.	.	.	36.53%	221300
227	203.12	233100	.	.	.	.
229	.	.	.	.	37.98%	331110

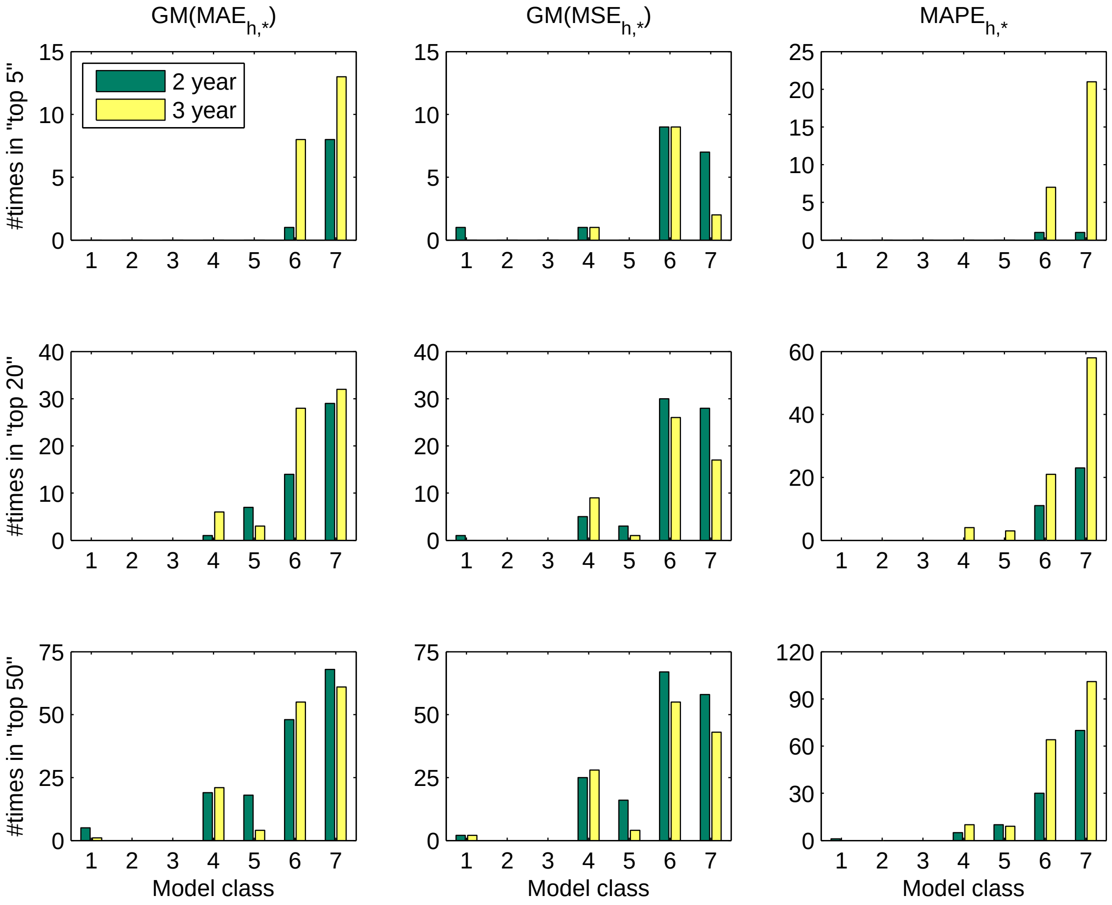

Figure 5: Histograms showing how many times models from a given family (**1***000*, **2*****00, **3******0, **4*****0*, **5*****00, **6******, **7*****0) are ranked in the top 5 (*top row*), top 20 (*center row*) and top 50 (*bottom row*) of all 300 models according to $GM(MAE_{h,*})$, $GM(MSE_{h,*})$ and $MAPE_{h,*}$ (*in columns, from left to right*) for each of the six forecast horizons $h = 1, ..., 6$. Clearly models from families 6***** and 7****0 dominate the rankings. Note the different scale in the right column (i.e., for $MAPE_{h,*}$).

families **6******* and **7****0**. Models from the remaining five families are generally much further down the list, see the bottom rows in Table 3. The best models from families 4***0* and 5***00 are respectively ranked at no. 55 and 80 for GM(MAE$_{*,*}$), 16 and 79 for GM(MSE$_{*,*}$) and 68 and 72 for MAPE$_{*,*}$. Next in line are the simple models (1*000*) – no. 87, 110 and 137 for the three error measures, respectively. Finally, the sine based models (2***00 and 3****0) close the list with ranks below 200, i.e., they are classified among the 33% worst performing models.

The best five models in terms of each global measure are emphasized in bold with an index indicating their rank: ^{a,b,c,d,e} for GM(MAE$_{*,*}$), ^{1,2,3,4,5} for GM(MSE$_{*,*}$) and ^{A,B,C,D,E} for MAPE$_{*,*}$. Three out of the top five models according to GM(MAE$_{*,*}$) are from the **73*110** subfamily, i.e., wavelets with a linear decay to the median, calibrated on a three-year window using approximation S_6 and with spikes replaced by the mean of the deseasonalized prices; they only differ in the choice of the wavelet family. The top four models according to MAPE$_{*,*}$ are from the same small subfamily (73*110), while the top 17 models are from the **73***0** subfamily. Clearly the three-year calibration window leads to better forecasts as measured by MAPE$_{*,*}$. On the other hand, the ranking according to GM(MSE$_{*,*}$) is dominated by 6***** models. Four out of the top five models models are from the **62*322** subfamily, i.e., wavelets with an exponential decay to the median with the decay parameter $\lambda = \frac{1}{180}$, calibrated on a two-year window using approximation S_8 and with spikes replaced by the upper/lower 2.5% quantiles of the deseasonalized prices; they only differ in the choice of the wavelet family.

Generally models 6***** with an exponential decay to the median perform better if the spikes are replaced by the upper/lower 2.5% quantiles of the deseasonalized prices (models with ‘2’ as the 5th digit), while models 7****0 perform better if the spikes are replaced by the mean of the deseasonalized prices (with ‘1’ as the 5th digit). This may indicate that for the former the exponential decay is too fast (even for the small decay parameter $\lambda = \frac{1}{180}$) and has to start at a more extreme level, while for models with a linear decay to the median the decay is too slow and the decay should start at a more typical, less extreme level.

These observations are confirmed by the results presented in Figure 5 where we plot histograms showing how many times models from a given family are ranked in the top 5 (which roughly corresponds to top 2%), top 20 (or 7%) and top 50 (or 17%) of all 300 models according to $GM(MAE_{h,*})$, $GM(MSE_{h,*})$ and $MAPE_{h,*}$ for each of the six forecast horizons $h = 1, ..., 6$. To be more precise, the ‘top 5’ histogram for $GM(MAE_{h,*})$ is constructed based on the five best models for the first forecast horizon and all six datasets, i.e., according to $GM(MAE_{1,*})$, the five best models for the second forecast horizon and all six datasets, i.e., according to $GM(MAE_{2,*})$, etc. In total $5 \times 6 = 30$ models are considered. Note that the models do not have to be unique as some of them may be ranked in the ‘top 5’ for more than one forecast horizon. For instance, model 732110 is 4th according to $GM(MAE_{4,*})$ and 2nd according to $GM(MAE_{5,*})$, see Table 5.

Like in Table 3, also in Figure 5 models from families 6***** and 7****0 dominate the rankings – with family 6***** performing better in terms of $GM(MSE_{h,*})$ and family 7****0 in terms of $GM(MAE_{h,*})$ and $MAPE_{h,*}$. On the other hand, not a single sine-based model (families 2***00 and 3****0) is ranked in the ‘top 50’. Note also that in terms of $MAPE_{h,*}$ the three-year calibration window is preferred, while for the other two error measures the evidence is not that clear.

5.2. Performance across the forecasting horizons

The performance of the models is not uniform across the forecasting horizons, see Tables 4 and 5 where we list the top five models over all six datasets according to the three error measures: GM(MAE$_{h,*}$), GM(MSE$_{h,*}$) and MAPE$_{h,*}$ for $h = 1$ to 6. For instance, the top four models according to GM(MSE$_{*,*}$) make it to the ‘top 5’ lists for the intermediate forecasting horizons of 31 to 182 days, but none of them makes it to the ‘top 5’ lists for the very short-term horizon (1-7 days) nor the long-term horizons (183-365 days). On the other hand, the top four models according to MAPE$_{*,*}$ are listed in the ‘top 5’ rankings for horizons of 91 to 274 days (i.e., the second and the third quarter of the one-year forecasting window), but none of them are listed in the ‘top 5’ list for the very short-term horizon (1-7 days) nor the very long-term horizon (275-365 days). Apparently the very short and the very long end of the one year forward curve requires other models. Only one of the emphasized in Table 3 models, i.e. **633122**, is listed in the 'top 5' for the shortest forecasting horizon. This model also performs very well for the second horizon (8-30 days), but not for any of the longer horizons. In contrast to all other highly ranked in Table 3 models from family 6*****, this model is based on the more sensitive to price changes approximation S_6 (i.e., with '1' as the 4th digit), rather than on the more smooth approximation S_8 (with '3' as the 4th digit). This increased sensitivity to local price fluctuations is a common feature of all 'top 5' ranked models for the very short term horizon, see the upper part of Table 4.

It also seems that the exponential decay to the median is too fast for the 3rd quarter of the forecast year (despite the small decay parameter $\lambda = \frac{1}{180}$), but much better than the linear decay for the horizon of 31 to 90 days. Finally, if we were to look for one 'best performing model' then only one of the emphasized in Table 3 models, i.e. **733110**, could be found in the 'top 5' lists for four out of the six forecasting horizons. This model is also globally ranked best or second best with respect to GM(MAE$_{*,*}$) and MAPE$_{*,*}$. It is a member of the well performing 73*110 subfamily, i.e., it is a Coiflets wavelet of order 2 ('coif2') with a linear decay to the median, calibrated on a three-year window using approximation S_6 and with spikes replaced by the mean of the deseasonalized prices. The best model from family 6***** in this competition is **631312**, i.e, a Daubechies wavelet of order 12 ('db12') with an exponential decay to the median with the decay parameter $\lambda = \frac{1}{180}$, calibrated on a three-year window using approximation S_8 and with spikes replaced by the mean of the deseasonalized prices. It is one of only two 6***1* models in the 'top 20' lists in Table 3. The substitution of spikes by the mean of the deseasonalized prices ('1' as the fifth digit) tends to make the price forecast less extreme for the intermediate time horizons.

For the forecasting horizon of 1 to 7 days only models from the 6**12* and 7**12* subfamilies are listed in Table 4, i.e., wavelets with an exponential or a linear decay to the median, calibrated using the more sensitive approximation S_6 and with spikes replaced by the upper/lower 2.5% quantiles of the deseasonalized prices. This outcome for the short end of the forward curve could have been expected – the more sensitive approximation level allows for a better local fit and, hence, a better short term forecast. On the other hand, for the forecasting horizon of 275 to 365 days (i.e., the '4th quarter') all but one model are from family 7****0, see Table 5. The exception is a simple model – 120001 with the mean price in the calibration window as the forecast – which is ranked poorly according to the linear error measures, but very good (2nd) with respect to GM(MSE$_{6,*}$).

5.3. Significance of the results

Finally, we may ask how significant are the differences between the models. In particular, does the domination of models from families 6***** and 7****0 observed in Tables 3-5 mean that models from the other families are inferior? To check this we performed a Friedman test to examine significant differences between the forecasting performance of selected models. The Friedman test is a nonparametric version of the classical two-way analysis of variance (ANOVA), and tests the null hypothesis that all matched samples are drawn from the same population, or equivalently, from different populations with the same distribution (Hochberg and Tamhane, 1987; Sprent and Smeeton, 2001). Unlike a classical ANOVA, the test does not require the assumption that all samples come from a population with a normal distribution. Examining the distribution of the forecasting errors, the assumption of normality for the population would not be justified.

Table 4: Top five models according to the three error measures and over all six datasets. The models are ranked with respect to MAPE$_{h,*}$, independently for each of the three shorter forecasting horizons: 1-7 days, 8-30 days and 31-90 days. The best five models in terms of each of the global error measures, i.e., GM(MAE$_{*,*}$), GM(MSE$_{*,*}$), MAPE$_{*,*}$), are emphasized in bold; the index indicates their rank, see Table 3. Additionally, the only three models in Tables 4-5 not belonging to families 6***** or 7****0 are marked with a dagger (‘†’).

Forecasting horizon 1-7 days

Model	GM(MAE$_{1,*}$)	rank	GM(MSE$_{1,*}$)	rank	**MAPE**$_{1,*}$	**rank**
734120	9.14	1	16.02	3	16.08%	1
634122	12.09	3	16.54	5	16.09%	2
633122^{d}	28.76	17	21.55	12	16.20%	3
732120	14.20	4	14.34	1	16.21%	4
731120	10.63	2	19.16	9	16.21%	5
631122	14.39	5	16.65	6	16.26%	8
721120	21.65	11	14.68	2	16.30%	10
621122	27.91	15	16.49	4	16.35%	12

Forecasting horizon 8-30 days

Model	GM(MAE$_{2,*}$)	rank	GM(MSE$_{2,*}$)	rank	**MAPE**$_{2,*}$	**rank**
633112	40.99	10	60.50	53	20.26%	1
733110^{a,B}	40.07	8	70.99	73	20.29%	2
633122^{d}	26.03	1	27.95	3	20.42%	3
733120	32.59	4	50.32	30	20.61%	4
632112	59.54	56	82.61	98	20.68%	5
634222	40.35	9	31.19	5	20.88%	17
731310^{b,E}	30.87	3	34.90	7	21.05%	38
632222	35.43	5	38.85	14	21.05%	39
631222	30.32	2	35.47	10	21.10%	43
631312^{4}	37.25	6	26.97	2	21.30%	62
631322^{5}	38.53	7	18.68	1	21.40%	69
431302†	43.38	15	30.10	4	21.80%	95

Forecasting horizon 31-90 days

Model	GM(MAE$_{3,*}$)	rank	GM(MSE$_{3,*}$)	rank	**MAPE**$_{3,*}$	**rank**
731310^{b,E}	20.44	2	26.33	11	24.96%	1
631312^{4}	18.01	1	14.49	1	25.00%	2
623322^{3}	22.78	4	18.36	3	25.20%	3
631322^{5}	21.47	3	15.27	2	25.21%	4
733110^{a,B}	29.68	9	58.82	53	25.36%	5
624322^{2}	29.72	10	21.65	5	25.45%	10
622322^{1}	37.90	20	20.58	4	25.58%	19
631222	24.88	5	32.14	16	25.64%	25

Table 5: Top five models according to the three error measures and over all six datasets. The models are ranked with respect to MAPE$_{h,*}$, independently for each of the three longer forecasting horizons: 91-182 days, 182-274 days and 275-365 days. The best five models in terms of each of the global error measures, i.e., GM(MAE$_{*,*}$), GM(MSE$_{*,*}$), MAPE$_{*,*}$), are emphasized in bold; the index indicates their rank, see Table 3. Additionally, the only three models in Tables 4-5 not belonging to families 6***** or 7****0 are marked with a dagger (‘†’).

Forecasting horizon 91-182 days (2nd quarter)						
Model	GM(MAE$_{4,*}$)	rank	GM(MSE$_{4,*}$)	rank	**MAPE**$_{4,*}$	**rank**
734110^{c,A}	12.11	1	24.63	11	28.71%	1
731310^{b,E}	17.96	2	18.92	7	28.71%	2
731110^{D}	18.22	3	27.17	16	28.74%	3
732110^{e,C}	18.91	4	28.65	17	28.77%	4
733110^{a,B}	20.46	5	34.97	25	28.83%	5
631312^{4}	23.46	7	17.87	4	28.97%	7
632322	27.81	11	17.68	3	29.13%	9
624322^{2}	26.57	10	12.86	1	29.46%	19
623322^{3}	27.88	13	18.03	5	29.55%	27
622322^{1}	33.23	21	16.10	2	29.70%	43
Forecasting horizon 183-274 days (3rd quarter)						
Model	GM(MAE$_{5,*}$)	rank	GM(MSE$_{5,*}$)	rank	**MAPE**$_{5,*}$	**rank**
734110^{c,A}	23.87	7	42.80	24	31.84%	1
732110^{e,C}	19.83	2	39.44	17	31.87%	2
731110^{D}	25.28	8	43.11	27	31.89%	3
733110^{a,B}	22.47	4	45.31	34	31.94%	4
731120	37.87	28	63.47	60	32.16%	5
722110	19.74	1	34.11	9	32.33%	17
721110	22.82	5	34.44	11	32.36%	18
721310	23.02	6	22.29	4	32.50%	21
721320	19.97	3	24.46	5	32.51%	23
621322	29.35	15	10.49	1	32.56%	27
421302†	43.61	36	16.26	2	32.83%	57
621312	39.32	32	20.42	3	32.86%	62
Forecasting horizon 275-365 days (4th quarter)						
Model	GM(MAE$_{6,*}$)	rank	GM(MSE$_{6,*}$)	rank	**MAPE**$_{6,*}$	**rank**
732220	98.36	110	114.47	125	33.14%	1
734220	105.44	122	117.99	132	33.17%	2
723220	14.67	1	29.13	5	33.17%	3
733220	107.36	129	122.03	140	33.18%	4
734120	108.34	131	126.94	149	33.18%	5
724220	17.24	3	22.11	3	33.18%	9
724210	24.49	4	35.87	8	33.19%	11
723210	25.14	5	45.70	19	33.20%	16
721220	25.98	6	27.39	4	33.23%	22
722220	16.59	2	18.98	1	33.23%	23
120001†	74.27	66	19.02	2	36.68%	225

Table 6: The table provides results for the multiple comparison procedure using Tukey's HSD criterion, indicating for each of the models which of the other models performs significantly worse / significantly better for the considered criterion, at the $\alpha = 0.05$ significance level. For each forecasting horizon $h = 1, ..., 6$ the results are based on bootstrapped subsamples of 1000 MAPE$_{h,*}$ errors. MAPE$_{h,*}$ errors based on the whole sample are also reported; the lowest for each forecasting horizon is emphasized in bold. The seven models used in the comparison were selected as those performing best in each model family (**1***000*, **2*****00, **3******0, **4*****0*, **5*****00, **6*******, **7******0) with respect to the global measure MAPE$_{*,*}$, see the last two columns in Table 3.

Model	MAPE$_{1,*}$	Worse / Better	MAPE$_{2,*}$	Worse / Better	MAPE$_{3,*}$	Worse / Better
120005	19.59%	{2,3} / –	24.30%	{2,3} / {5,6,7}	28.10%	{2,3} / {4,5,6,7}
221300	29.58%	– / All	31.09%	– / {1,4,5,6,7}	34.52%	– / {1,4,5,6,7}
331110	26.35%	{2} / {1,4,5,6,7}	28.03%	– / {1,4,5,6,7}	31.67%	– / {1,4,5,6,7}
431302	18.74%	{2,3} / {5,7}	21.80%	{2,3} / –	25.55%	{1,2,3} / –
531300	18.51%	{2,3,4} / –	22.01%	{1,2,3} / –	26.26%	{1,2,3} / –
631322^{5}	18.59%	{2,3} / –	21.40%	{1,2,3} / –	**25.21%**	{1,2,3,7} / –
734110^{c,A}	**16.84%**	{2,3,4} / –	**20.82%**	{1,2,3} / –	25.38%	{1,2,3} / {6}
Model	MAPE$_{4,*}$	Worse / Better	MAPE$_{5,*}$	Worse / Better	MAPE$_{6,*}$	Worse / Better
120005	31.53%	{2,3} / {4,5,6,7}	33.42%	{2,3} / {7}	33.35%	{2,3,4,6} / –
221300	36.91%	– / {1,4,5,6,7}	37.61%	– / {1,4,5,6,7}	38.28%	{3} / {1,5,7}
331110	35.20%	– / {1,4,5,6,7}	43.98%	– / {1,4,5,6,7}	42.30%	– / All
431302	29.44%	{1,2,3} / –	33.27%	{2,3} / {5,7}	34.26%	{3} / {1,5,7}
531300	29.91%	{1,2,3} / –	33.15%	{2,3,4,6} / –	33.65%	{2,3,4,6} / –
631322^{5}	29.06%	{1,2,3} / –	32.85%	{2,3} / {5,7}	33.75%	{3} / {1,5,7}
734110^{c,A}	**28.71%**	{1,2,3} / –	**31.84%**	{1,2,3,4,6} / –	**33.26%**	{2,3,4,6} / –

Furthermore, both ANOVA and the Friedman test make an assumption of independence. And this clearly is not met by the model errors in our empirical study as we use a rolling window scheme. To cope with this and break the dependence structure, we used bootstrapped subsamples of 1000 errors instead of the full samples of 8580 errors; recall from Section 4.1 that for all six datasets we have 8580 forecasting windows in total. The bootstrapped subsamples contained matched errors, i.e., for each forecasting horizon ($h = 1, ..., 6$) errors for the same randomly chosen 1000 forecasting windows were selected for each model. Seven best performing models in terms of MAPE$_{*,*}$ were selected for the significance test – one from each model family (**1***000*, **2*****00, **3******0, **4*****0*, **5*****00, **6*******, **7******0), see the last two columns in Table 3. For instance, model 734110 is the best according to MAPE$_{*,*}$ and third best according to GM(MAE$_{*,*}$) while model 631322 is fifth best according to GM(MSE$_{*,*}$) and the best of all 6***** family models in terms of MAPE$_{*,*}$.

Rejecting the null of the Friedman test only provides statistical evidence for at least one of the samples being from a population with a different distribution. However, the test does not provide detailed information on which of the samples are significantly different. A test that can do so is called a multiple comparison procedure. In this study we use Tukey's honestly (or wholly) significant difference test (Tukey's HSD or Tukey's WSD) which is optimal for the comparison of groups with equal sample sizes (Hochberg and Tamhane, 1987; Maxwell and Delaney, 2004). For each of the seven models Table 6 indicates which of the other models perform significantly worse or significantly better in terms of $\text{MAPE}_{h,*}$, at the $\alpha = 0.05$ significance level and independently for each forecasting horizon $h = 1, ..., 6$. Note that when all other models were significantly better (or worse) than a particular model this is indicated by 'All', while '–' indicates that none of the other models provided significantly better (or worse) results. Being based on a nonparametric test statistic, it can be expected that the multiple comparison procedure will not be able to distinguish significant differences between all of the considered models. However, many of the differences between the models observed earlier in this Section are identified as significant.

The best sine-based models (221300 and 331110) perform significantly worse than the best simple model (120005) and the best wavelet-based models (431302, 531300, 631322 and 734110) across all six forecasting horizons; except for model 221300 in the most distant horizon $h = 6$ (i.e., 275-365 days ahead) which is not found to be significantly worse than the spike unfiltered wavelet-based models 431302 and 531300. The simple model 120005 performs surprisingly well for the closest ($h = 1$) and the two most distant ($h = 5, 6$) forecasting horizons. Interestingly, if we repeat the analysis but take the worst performing models in terms of $\text{MAPE}_{*,*}$ from the simple and wavelet-based families (namely, models 120002, 423101, 523100, 622111 and 721220) and compare them with the best sine-based models we obtain that all four wavelet-based models perform significantly better across all six forecasting horizons than the remaining three models (120002, 221300 and 331110)! On the other hand, the worst simple model (120002 – linear regression of the spot price in the two-year calibration window extrapolated into the forecasting window) performs much worse than the best simple model (120005) and is generally comparable to the best sine-based models.

Furthermore, models 531300 and 734110 (i.e., wavelets with a linear decay to the median) perform significantly better than models 431302 and 631322 (i.e., wavelets with an exponential decay to the median) for the two longer horizons of 183-365 days ahead ($h = 5, 6$). However, model 631322 is significantly better than model 734110 for horizon $h = 3$ (i.e., 31-90 days) and comparable to it for horizons $h = 1, 2$ and 4. Model 531300 is the only one that has no significantly better competitors for all six forecasting horizons, however, model 734110 yields lower $\text{MAPE}_{h,*}$ for all $h = 1, ..., 6$ and has a larger number of competing models performing significantly worse for horizon $h = 5$. Overall we can conclude that model 734110 is the best performing model, with model 631322 trailing closely by. Moreover, any of the wavelet-based models is better than the sine-based models.

6. Conclusions

In this paper we have presented the results of a thorough study on estimation and forecasting of the long-term seasonal component (LTSC) of electricity spot prices. We have considered a battery of models:

- 12 simple linear models, including models with a deterministic function linearly or exponentially decaying from the last observed spot price to the median in the forecasting window (model family 1*000*),

- 24 sine-based models fitted to raw prices (2***00),

- 48 sine-based models fitted to spike-filtered prices (3****0),
- 48 wavelet-based models fitted to raw prices and a function exponentially decaying to the median in the forecasting window (4***0*),
- 24 wavelet-based models fitted to raw prices and a function linearly decaying to the median in the forecasting window (5***00),
- 96 wavelet-based models fitted to spike-filtered prices and a function exponentially decaying to the median in the forecasting window (6*****),
- 48 wavelet-based models fitted to spike-filtered prices and a function linearly decaying to the median in the forecasting window (7****0).

The models differ in the length of the calibration window, the number and the periods of the sine functions, the wavelet families and approximation levels, etc. For details see Tables 1-2.

Using daily baseload spot prices from six major power markets – NSW in Australia, EEX and Nord Pool in Europe, NEPOOL, NYISO and PJM in the U.S. – we find that wavelet-based models (families 4***0*, 5***00, 6***** and 7****0) are better in terms of forecasting spot prices up to a year ahead than sine-based models (families 2***00 and 3****0). This observation is valid for all three error measures (MAE, MSE, MAPE) both globally over all forecasting horizons (see Table 3) as well as individually across the six forecasting horizons (see Tables 4-5 and Figure 5). The statistical significance of this finding is confirmed in Table 6 using MAPE errors and Tukey's HSD multiple comparison test.

This result questions the validity and usefulness of stochastic models of spot electricity prices built on sinusoidal long-term seasonal components. It also gives a clear recommendation for using wavelet-based models for estimating and forecasting the LTSC. Not only are these models able to provide a good in-sample fit in the calibration window (and generally much better than that of sine-based models with a reasonable number of sine functions), but also yield significantly better forecasts up to a year ahead.

Overall we can conclude that model 734110 (i.e., a Coiflets wavelet of order 4 with a linear decay to the median, calibrated on a three-year window using approximation S_6 and with spikes replaced by the mean of the deseasonalized prices) is the best performing model, with model 631322 (i.e., a Daubechies wavelet of order 12 with an exponential decay to the median with the decay parameter $\lambda = \frac{1}{180}$, calibrated on a three-year window using approximation S_8 and with spikes replaced by the upper/lower 2.5% quantiles of the deseasonalized prices) trailing closely by. However, as the results reported in Section 5.3 indicate, the choice of the wavelet family – Coiflets or Daubechies – is not critical. Nor is the choice of the remaining parameters, despite some subtle differences discussed in Section 5.

Surprisingly, some simple models (including 120005, i.e., a deterministic function exponentially decaying in the forecasting window from the last observed spot price to the median spot price in the two-year calibration window) perform very well for the closest (1-7 days) and the two most distant (183-365 days) forecasting horizons. On the other hand, some simple models (like 120002, i.e., linear regression of the spot price in the two-year calibration window extrapolated into the forecasting window) perform much worse and generally comparable over all forecasting horizons to the poorly performing sine-based models. Unfortunately, the better performing simple models are discontinuous (on the day the forecast is made) and, hence, can be used for forecasting the spot price up to a year ahead but should not be used for deseasonalizing the electricity spot price series before fitting the stochastic model.

Finally, let us comment on two alternative approaches to modeling and forecasting the LTSC which were briefly mentioned by Janczura and Weron (2010, 2012) in the context of wavelet-based models. The first is to use forward looking information, like smoothed forward electricity curves (Benth et al., 2007; Borak and Weron, 2008). While this is a potentially promising approach, it has to be taken into account that forward prices include risk premia, which should somehow be separated from the spot price forecast for it to be useful. And this is not an easy task since risk premia vary over time (Botterud et al., 2010; Huisman and Kilic, 2012; Weron, 2008). There are also some discouraging examples. For instance, Stevenson et al. (2006) used consensus forecasts of wholesale electricity spot prices issued by the Australian Financial Market Association (AFMA). Given the lack of liquidity in electricity derivative contracts traded on the Sydney Futures Exchange (SFE) at the time of their study, the accepted market forward price was the AFMA price rather than a market traded price. As Stevenson et al. report, the AFMA (forward) prices turned out to be misleading, strongly biased estimates of the future spot price. In a related study Redl et al. (2009) observe that also in the German EEX and the Scandinavian Nord Pool markets differences between forward prices in the trading period and spot prices in the delivery period are significant. They further note that trading strategies of market participants seem to rely heavily on current spot prices instead of fundamental modeling approaches. These results question the predictive power of forward prices. The second alternative approach is to utilize fundamental information like weather (Jabłońska et al., 2011), generation and demand (Cartea et al., 2009) or inventory levels (Douglas and Popova, 2008). There are, however, some problems related with this. The most important one being the limited availability of good quality forecasts of these fundamental factors for longer time horizons. In particular, Redl et al. (2009) provide evidence for misjudgment of future fundamental generation and demand conditions by market participants.

Taking all this into account we may conclude that the wavelet-based LTSC models considered in this study provide relatively simple and accurate means of describing and forecasting the LTSC of spot electricity prices. The fact that these models are calibrated to publicly available historical data makes them an attractive building block of stochastic models of spot electricity prices.

Acknowledgements

This paper has benefited from conversations with the participants of the Energy Finance Christmas Workshop (EFC11), the Macquarie Risk Day 2012, the European Energy Market (EEM12) conference, the Conference on Energy Finance EF2012 and the seminars at National University of Singapore, Adam Mickiewicz University in Poznań and Wrocław University of Technology. This work was supported by funds from the National Science Centre (NCN, Poland) through grant no. 2011/01/B/HS4/01077.

References

Becker, R., Hurn, S., Pavlov, V. (2007) Modelling spikes in electricity prices. The Economic Record 83(263), 371-382.

Benth, F.E., Kiesel, R., Nazarova, A. (2012). A critical empirical study of three electricity spot price models. Energy Economics 34, 1589-1616.

Benth, F. E., Koekebakker, S., Ollmar, F. (2007) Extracting and applying smooth forward curves from average-based commodity contracts with seasonal variation. Journal of Derivatives – Fall, 52-66.

Benz, E., Trück, S. (2006) CO_2 emission allowances trading in Europe – Specifying a new class of assets. Problems and Perspectives in Management 4(3), 30-40.

Bhanot, K. (2000) Behavior of power prices: Implications for the valuation and hedging of financial contracts. The Journal of Risk 2, 43-62.

Bierbrauer, M., Menn, C., Rachev, S.T., Trück, S. (2007) Spot and derivative pricing in the EEX power market. Journal of Banking and Finance 31, 3462-3485.

Borak, S., Weron, R. (2008) A semiparametric factor model for electricity forward curve dynamics. Journal of Energy Markets 1(3), 3-16.

Bordignon, S. et al. (2012). Combining day-ahead forecasts for British electricity prices. Energy Economics, forthcoming.

Botterud, A., Kristiansen, T., Ilic, M.D. (2010) The relationship between spot and futures prices in the Nord Pool electricity market. Energy Economics 32, 967-978.

Cartea, A., Figueroa, M. (2005) Pricing in electricity markets: A mean reverting jump diffusion model with seasonality. Applied Mathematical Finance 12(4), 313-335.

Cartea, A., Figueroa, M., Geman, H. (2009) Modelling Electricity Prices with Forward Looking Capacity Constraints. Applied Mathematical Finance 16(2), 103-122.

Conejo, A.J., Plazas, M.A., Espínola, R., Molina, A.B. (2005). Day-ahead electricity price forecasting using the wavelet transform and ARIMA models. IEEE Transactions on Power Systems 20(2), 1035-1042.

De Jong, C. (2006) The nature of power spikes: A regime-switch approach. Studies in Nonlinear Dynamics & Econometrics 10(3), Article 3.

Douglas, S., Popova, J. (2008) Storage and the electricity forward premium. Energy Economics 30, 1712-1727.

Erlwein, C., Benth, F.E., Mamon, R. (2010) HMM filtering and parameter estimation of an electricity spot price model. Energy Economics 32, 1034-1043.

Eydeland, A., Wolyniec, K. (2012) Energy and Power Risk Management (2nd ed.). Wiley, Hoboken, NJ.

Fanone, E., Gamba, A., Prokopczuk, M. (2012) The case of negative day-ahead electricity prices. Energy Economics, In press, doi:10.1016/j.eneco.2011.12.006.

Fleten, S.-E., Heggedal, A.M., Siddiqui, A. (2011) Transmission capacity between Norway and Germany: a real options analysis. Journal of Energy Markets 4(1), 121-147.

Fryźlewicz, P., van Bellegem, S., von Sachs, R. (2003) Forecasting non-stationary time series by wavelet process modelling. Annals of the Institute of Statistical Mathematics 55, 737-764.

Geman, H., Roncoroni, A. (2006) Understanding the fine structure of electricity prices. Journal of Business 79, 1225-1261.

Gianfreda, A., Grossi, L. (2012) Forecasting Italian electricity zonal prices with exogenous variables. Energy Economics 34, 2228-2239.

Haldrup, N., Nielsen, F.S., Nielsen, M.Ø. (2010) A vector autoregressive model for electricity prices subject to long memory and regime switching. Energy Economics 32, 1044-1058.

Hamilton, J. (2009) Causes and Consequences of the Oil Shock of 2007-08. Brookings Papers on Economic Activity 2009/1, 215-261.

Haugom, E., Ullrich, C.J. (2012) Forecasting spot price volatility using the short-term forward curve. Energy Economics 34, 1826-1833.

Härdle, W., Kerkyacharian, G., Picard, D., Tsybakov, A. (1998) Wavelets, Approximation and Statistical Applications. Lecture Notes in Statistics 129. Springer-Verlag, New York.

Higgs, H., Worthington, A. (2008) Stochastic price modeling of high volatility, mean-reverting, spike-prone commodities: The Australian wholesale spot electricity market. Energy Economics 30, 3172-3185.

Hochberg, Y., Tamhane, A.C. (1987) Multiple Comparison Procedures. Wiley, Hoboken, NJ.

Hollander, M., Wolfe, D.A. (1999) Nonparametric Statistical Methods. Wiley, Hoboken, NJ.

Huisman, R., Kilic, M. (2012) Electricity futures prices: Indirect storability, expectations, and risk premiums. Energy Economics 34, 892-898.

Jabłońska, M., Nampala, H., Kauranne, T. (2011) The multiple-mean-reversion jump-diffusion model for Nordic electricity spot prices. Journal of Energy Markets 4(2), 3-25.

Janczura, J., Weron, R. (2010). An empirical comparison of alternate regime-switching models for electricity spot prices. Energy Economics 32, 1059-1073.

Janczura, J., Weron, R. (2012). Efficient estimation of Markov regime-switching models: An application to electricity spot prices, AStA - Advances in Statistical Analysis 96(3), 385-407.

Janczura, J., Trüeck, S., Weron, R., Wolff, R. (2012). Identifying spikes and seasonal components in electricity spot price data: A guide to robust modeling, submitted. Working paper version available from MPRA: http://mpra.ub.uni-muenchen.de/39227/.

Kanamura, T., Ōhashi, K. (2008) On transition probabilities of regime switching in electricity prices. Energy Economics 30, 1158-1172.

Keles, D., Genoese, M., Möst, D., Fichtner, W. (2012a) Comparison of extended mean-reversion and time series models for electricity spot price simulation considering negative prices. Energy Economics 34, 1012-1032.

Keles, D., Hartel, R., Möst, D., Fichtner, W. (2012b) Compressed-air energy storage power plant investments under uncertain electricity prices: An evaluation of compressed-air energy storage plants in liberalized energy markets. Journal of Energy Markets 5(1), 53-84.

Knittel, C.R., Roberts, M.R. (2005) An empirical examination of restructured electricity prices. Energy Economics 27, 791-817.

Lucia, J.J., Schwartz, E.S. (2002) Electricity prices and power derivatives: Evidence from the Nordic Power Exchange. Review of Derivatives Research 5, 5-50.

Maxwell, S.E., Delaney, H.D. (2004) Designing Experiments and Analyzing Data, 2nd ed. Lawrence Erlbaum Associates, Inc.

Nomikos, N.K., Soldatos, O.A. (2010) Analysis of model implied volatility for jump diffusion models: Empirical evidence from the Nordpool market. Energy Economics 32, 302-312.

Pilipovic, D. (1998) Energy Risk: Valuing and Managing Energy Derivatives. McGraw-Hill, New York.

Percival, D.B., Walden, A.T. (2000) Wavelet Methods for Time Series Analysis. Cambridge University Press.

Redl, C., Haas, R., Huber, C., Böhm, B. (2009) Price formation in electricity forward markets and the relevance of systematic forecast errors. Energy Economics 31, 356-364.

Schlueter, S. (2010) A long-term/short-term model for daily electricity prices with dynamic volatility. Energy Economics 32, 1074-1081.

Schlueter, S., Deuschle, C. (2010) Using wavelets for time series forecasting – Does it pay off? IWQW Discussion Paper 4/2010.

Seifert, J., Uhrig-Homburg, M. (2007) Modelling jumps in electricity prices: theory and empirical evidence. Review of Derivatives Research 10, 59-85.

Shahidehpour, M., Yamin, H., Li, Z. (2002) Market Operations in Electric Power Systems: Forecasting, Scheduling, and Risk Management. Wiley.

Sprent, P., Smeeton, N.C. (2001) Applied nonparametric statistical methods, 3rd ed. CRC Press.

Stevenson, M. (2001) Filtering and forecasting spot electricity prices in the increasingly deregulated Australian electricity market. Research Paper No 63, Quantitative Finance Research Centre, University of Technology, Sydney.

Stevenson, M.J., Amaral, J.F.M., Peat, M. (2006) Risk management and the role of spot price predictions in the Australian retail electricity market. Studies in Nonlinear Dynamics and Econometrics 10(3), Article 4.

Trück, S., Weron, R., Wolff, R. (2007) Outlier treatment and robust approaches for modeling electricity spot prices. Proceedings of the 56th Session of the ISI. Available at MPRA: http://mpra.ub.uni-muenchen.de/4711/.

Weron, R. (2006) Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach. Wiley, Chichester.

Weron, R. (2008) Market price of risk implied by Asian-style electricity options and futures. Energy Economics 30, 1098-1115.

Weron, R. (2009) Heavy-tails and regime-switching in electricity prices. Mathematical Methods of Operations Research 69(3), 457-473.

Weron, R., Bierbrauer, M., Trück, S. (2004a) Modeling electricity prices: jump diffusion and regime switching. Physica A 336, 39-48.

Weron, R., Simonsen, I., Wilman, P. (2004b) Modeling highly volatile and seasonal markets: Evidence from the Nord Pool electricity market. In: The Application of Econophysics, H. Takayasu (ed.), Springer, Tokyo, 182-191.

Wong, H., Ip, W.-C., Xie, Z., Lui, X. (2003). Modelling and forecasting by wavelets, and the application to exchange rates. Journal of Applied Statistics 30(5), 537-553.

Yousefi, S., Weinreich, I., Reinarz, D. (2005). Wavelet-based prediction of oil prices. Chaos, Solitons and Fractals 25, 265-275.